HNEDTI: Prediction of drug-target interaction based on heterogeneous network embedding

Zhangli Lu
School of Computer Science and Engineering
Central South University
Changsha Hunan, China
luzhangli@csu.edu.cn

Yake Wang
School of Computer Science and Engineering
Central South University
Changsha Hunan, China
wangyk@csu.edu.cn

Min Zeng
School of Computer Science and Engineering
Central South University
Changsha Hunan, China
zengmin@csu.edu.cn

Min Li*
School of Computer Science and Engineering
Central South University
Changsha Hunan, China
limin@mail.csu.edu.cn

***Abstract*—Identifying drug-target interactions (DTIs) is an important task in drug discovery. Various computational models have been proposed to predict potential association between drugs and targets. However, it is still a great challenge to accurately predict the potential drug-target interactions with rare known drug-target interactions. In this work, we propose a heterogeneous network embedding model to predict drug-target interactions, called HNEDTI. Based on the assumption that similar drugs share similar patterns of relationships with target proteins, we integrate the drug-drug similarity network, target-target similarity network and known drug-target interactions into a heterogeneous network. HNEDTI can learn more accurate feature representation of drugs and targets by extract both local and global information of the heterogeneous network from different lengths of meta-paths. The low dimensional feature representation vectors of drugs and targets are applied to random forest model to predict whether the given drug-target pair has an interaction. The evaluation on four benchmark datasets (Enzyme, Ion Channel, GPCR and Nuclear Receptor) shows that our method HNEDTI outperforms the previous methods.**

***Index Terms*—drug-target interaction, network embedding, machine learning**

I. Introduction

Drug discovery is the main objective of pharmaceutical science. Over the past decades, drug discovery technology has developed rapidly, and various methods including genomics, proteomics, and systems biology have been widely used in the identification of drug-target interactions and the development of novel drugs [1]–[3]. However, the development of a new drug is still costly and inefficient. Each new molecular entity (NME) will cost $ 1.8 billion for average, and takes a long period to put into use [4]. Furthermore, the rate of successful drug development has decreased in recent years. According to the US Food and Drug Administrations (FDA) statistical data, only about 20 drugs have been approved by FDA as NMEs each year [5]. In the process of drug discovery, one of the key steps is to identify interactions between drugs and targets. Because drug-target interactions can facilitate choosing a specific compound for a particular protein in the target-based drug discovery [6]. Thus, it is of great signification to identify the potential drug-target interactions for existing or abandoned drugs.

Recently, various network-based computational methods have been proposed to predict the underlying drug-target associations by integrating some heterogeneous networks about drugs and targets. Yang et al. developed a computational algorithm for finding potential drug targets and multiple target optimal intervention (MTOI) solutions by systematically analyzing the disease states and the desired states in the disease network [7]. Based on the assumption that similar drugs often interact with similar targets, Chen et al. proposed a Random Walk with Restart based method on the Heterogeneous network (NRWRH) to infer potential DTIs on a large scale [8]. Cheng et al. developed three supervised inference methods to predict DTIs, namely drug-based similarity inference (DBSI), target-based similarity inference (TBSI) and network-based inference (NBI) [9]. Campillos et al. developed a measure for side-effect similarity and integrated drug side-effect similarity and chemical similarity to infer whether two drugs share a target [10].

The above computational methods, especially network-based methods, have been proved to be effective in the prediction of drug-target interactions. However, most network-based methods are pure local model and cannot discover global network information. While the global network information is more helpful for the prediction of drug and target candidates without known interactions. To address the limitation, we propose a heterogeneous network embedding framework, called HNEDTI, to predict drug-target association. We integrate the drug-drug similarity network, target-target similarity network and known drug-target interactions into a heterogeneous network. The random walk technique is applied to extract various meta-paths and sample training data. In the representation learning section, we design a neural network model to capture the rich relationship embedded in the drug-target heterogeneous network and learn a latent low dimensional space to represent these drugs and targets. Finally, the low dimensional representation of drugs and targets are used to feed into a random forest model to infer potential DTIs.

Compared to traditional network-based methods, HNEDTI can learn both local and global information of the heterogeneous network from different lengths meta-paths. Short meta-paths contain the local information of the network, while long

978-1-7281-1867-3/19/$31.00 ©2019 IEEE
978-1-7281-1867-3/19/$31.00 ©2019 IEEE

meta-paths contain the global information of the network. Most traditional network-based methods only use the local network information to infer potential DTIs. However, the advantage of the global network information over the local network information which has been demonstrated in much previous research [11], [12]. Compared to traditional machine learning-based methods, our method can learn a more accurate representation of drug and target. The evaluation on four benchmark datasets (Enzyme, Ion Channel, GPCR and Nuclear Receptor) demonstrate that our method outperforms other methods. Hence, we can conclude that HNEDTI can be successfully applied to predict potential DTIs.

II. Methods

A. Problem Definition

The purpose of drug-target interaction prediction is to predict the score of unknown relationship in the interaction matrix by using known drug-target interactions and similarity information. Given the set of drugs as $D = \{d_1, \cdots, d_m\}$ and the set of targets as $T = \{t_1, \cdots, t_n\}$, where the m and n are the number of drugs and targets respectively. We use $R \in R^{m \times n}$ to denote the drug-target interactions matrix. R is a binary matrix with entries $r_{i,j} = 1$ denoting that a drug d_i has been experimentally verified to interact with a target t_j, $r_{i,j} = 0$ otherwise, according to the known drug-target interaction information. The drug-drug similarity matrix $S^D \in R^{m \times m}$ contains the similarity between different drugs, for each element $S^D_{i,j}$ represents the similarity between drugs d_i and d_j. Similarly, the target-target similarity matrix $S^T \in R^{n \times n}$ contains the similarity between different targets.

B. Construction Of Drug-target Heterogeneous Network

Our DTIs prediction model is based on the assumption that similar drugs often interact with similar targets. Thus, we integrate the drug-target interaction matrix R, drug-drug similarity matrix S^D and target-target similarity matrix S^T to construct a drug-target heterogeneous network. We set two similarity threshold parameter α and β for drug similarity matrix and target similarity matrix respectively to filter the edges with low similarity. Finally, combine the known drug-target interactions, drug-drug edge and target-target edge with higher similarity into drug-target heterogeneous network. Let $G = \{V, E\}$ denote the heterogeneous network, V is the nodes set including drug nodes and target nodes, and E is the edge set including drug-drug, drug-target, target-drug, and target-target.

C. Heterogeneous Network Embedding Model

Recently, network embedding methods have been successfully applied in various fields [13]–[15]. The drug-target heterogeneous network contains both intra- and inter-information of drugs and targets. To learn the feature representation of drugs and targets, we design a specific heterogeneous network embedding (HNE) framework. It captures embedded information in the network structure by exploiting different types of information among nodes and can learn representations of nodes from both local and global aspects, and can be used as input to supervised machine learning algorithms. The HNE model consists of two phases: (1) Random walks and training data preparation; and (2) Representation learning.

1) Random Walks And Training Data Preparation: To sample the relationship between all the nodes in the heterogeneous network $G = \{V, E\}$, random walk is employed to generate walk sequences. Formally, given the number of walks k and walk length l, we simulate a walk sequence $S = \{s_1, s_2, \cdots, s_{l-1}, s_l\}$ of fixed length l, s_i represent the i-th node in the walk sequence. The next node s_{i+1} generated by the following normalized probability distribution:

$$P(s_{i+1} = v | s_i = u) = \begin{cases} \frac{W_{u,v}}{Z}, & if (u,v) \in E \\ 0, & otherwise \end{cases} \quad (1)$$

where E is the edge set of heterogeneous network, $w_{u,v}$ is the weight of edge (u, v), and Z is the normalizing constant which is the sum of the weights of the all edges connected node u. The random walk starts from each node in the node sets V, and repeats walking k times in order to obtain a stable distribution. In the process of the random walk, we do not need to consider the node types and specify the meta-path.

In the training data preparation phase, we extract the training data from the $|V| \times k$ walk sequences generated by the random walk technique. We use D denotes the type of drug and T denotes the type of target, and call the sequence of node type meta-path. Let M denotes the set of all meta-paths, $|M|$ denotes the number of meta-paths. To train the HNE model, training data need to be prepared a triple $(x_1, x_2, M^{'})$ for $|M|$ prediction tasks corresponding to M. x_1 and x_2 represent two input nodes encoded with one-hot vector, $M^{'}$ is a binary array with entries $M^{'}_i = 1$ denoting that node x_1 and node x_2 have the corresponding meta-path, otherwise: $M^{'}_i = 0$.

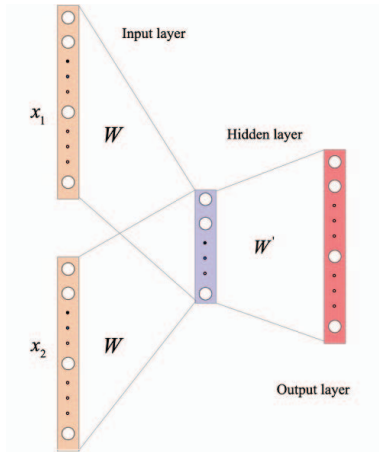

Fig. 1. The HNEDTI neural network model.

2) Representation Learning: As mentioned above, the representation learning of HNE model is a multiple prediction task, one for each meta-path. The model is a neural network with 3 layers feedforward. As shown in Figure 1, the input layer takes two one-hot vectors $\overrightarrow{x_1}$ and $\overrightarrow{x_2}$, which represents

input node x_1 and x_2 by $|V|$-dimension vectors. In the hidden layer, $\overrightarrow{x_1}$ and $\overrightarrow{x_2}$ are transformed into latent vectors $W\overrightarrow{x_1}$ and $W\overrightarrow{x_2}$, where W is a $|V| \times d$ weight matrix of hidden layer as well as the vector representation of all node, and d is the dimensionality of the hidden layer as well as the dimensionality of the vector representation. The hidden layer vector h is the Hadamard product of two latent vectors $W\overrightarrow{x_1}$ and $W\overrightarrow{x_2}$.

$$h = W\overrightarrow{x_1} \odot W\overrightarrow{x_2} \tag{2}$$

In the output layer, the model outputs a probability vector $\tilde{y}$ of length $|M|$ as follows:

$$\tilde{y} = \left[P(r = M_1|x_1, x_2), \cdots, P(r = M_{|M|}|x_1, x_2)\right] \tag{3}$$

Each item of $\tilde{y}$ represents the SoftMax probability that the input nodes x_1 and x_2 contain corresponding meta-path. The SoftMax probability is defined as follows:

$$P(r = M_i|x_1, x_2) = \frac{exp((W^{'}h)_i)}{\sum_{j=1}^{|M|} exp((W^{'}h)_j)} \tag{4}$$

where $W^{'}$ is a $|M| \times d$ matrix. Cross entropy is the loss function of the model. The objective of optimization is to minimize the loss function. Loss function can be defined as follows:

$$L(y, \tilde{y}) = -\sum_{i=1}^{|M|} y_i log\tilde{y}_i \tag{5}$$

D. Drug-target Interaction Prediction

In the representation learning phase, all the nodes in the heterogeneous network including drugs and targets are embedded into an embedding matrix W, which is a $|V| \times d$ matrix, where $|V|$ is the number of nodes and d is the dimension of vector representation. In association prediction section, we use the embedded vector of drugs and targets to predict the association between them. Given a drug d_i and a target t_j, let W_{d_i} and W_{t_j} denote their embedded vector, $(W_{d_i})^{'}$ and $(W_{t_j})^{'}$ are two normalized vectors of W_{d_i} and W_{t_j}, respectively. The vector representation of drug d_i and target t_j pair $P_{i,j} = (W_{d_i})^{'} \odot (W_{t_j})^{'}$, which is the Hadamard product of $(W_{d_i})^{'}$ and $(W_{t_j})^{'}$.

The random forest model is employed to build a binary classifier for DTI prediction. The input of the model is the vector representation of the drug-target pair, the corresponding association is the sample label. Finally, the model output a probability value indicates the possibility of the interaction corresponding to this drug target pair.

III. Evaluation And Results

A. Datasets

In this study, we use four benchmark drug-target interaction network in human including Enzyme (E), Ion Channel (IC), G-Protein Coupled Receptors (GPCR) and Nuclear Receptor (NR) [16]. Each of the four datasets contains a drug-target interaction matrix, a drug-drug similarity matrix and a target-target similarity matrix. Each entry of the drug-target interaction matrix indicates whether the interaction between the corresponding drug and the target is known or not. The interaction information is obtained from KEGG BRITE [17], BRENDA [18], SuperTarget [19], and DrugBank [20] databases. Drug-to-drug similarity scores are calculated by the SIMCOMP [21] based on chemical structures of compounds. The SIMCOMP method has been widely applied to calculate chemical structural similarity, especially the prediction of drug-target interactions [5], [6], [8], [11]. Target-to-target similarity scores are calculated by the normalized Smith-Waterman scores [22] based on amino acid sequences of target proteins from the KEGG GENES databases.

B. Comparison With Other Competing Methods

We evaluate HNEDTI method on four benchmark datasets: Enzyme, IC, GPCR and NR and compare it with five competing methods: BLM-NII [23], WNN-GIP [24], NetLapRLS [25], CMF [26] and BRDTI [27]. The performance of our method is evaluated via five times repeated 10-fold cross-validation as well as previous studies. We randomly assign the known DTIs from datasets to one out of ten splits, for each fold, a different split is used as a test set of the predictive model, while the remaining splits are used to construct the drug-target heterogeneous network and as the training set of predictive model. Then, we run five times repeated predictive model to obtain the final results. We adopt Area Under the Curve (AUC) as an evaluation metric.

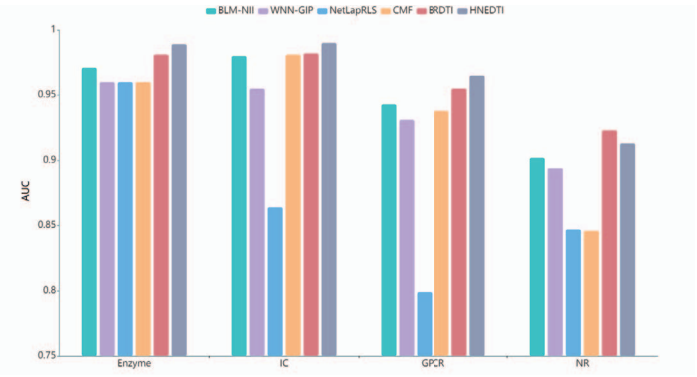

Fig. 2. The AUC scores of HNEDTI and other compared methods on four gold standard DTI datasets.

Figure 2 shows the AUC of HNEDTI compared with other methods on four gold standard datasets. HNEDTI method achieves the best result on Enzyme, Ion Channel and GPCR datasets and second-best on Nuclear Receptors dataset. The Nuclear Receptors dataset is too small with only 54 drugs, 26 targets and 90 known interactions, the training samples are not sufficient for HNEDTI to learn more accurate representation. So, we can conclude that HNEDTI significantly outperforms other methods and can be effectively applied to infer drug-target interactions.

IV. Discussions And Conclusions

Prediction of drug-target interactions plays an important role in drug repositioning and biological processes. The relationships between drugs and targets provide lots of information to identify new candidate associations. In this study, we proposed a heterogeneous network embedding model HNEDTI. Compared to the traditional network-based methods, HNEDTI can capture local and global information of the heterogeneous network, and is more powerful to learn the representation of drugs and targets.

To evaluate the performance of our model, we carry out experiments on four benchmark datasets (Enzyme, Ion Channel, GPCR and Nuclear Receptor). The result of AUC score via five times repeated 10-fold cross-validation outperforms other methods including BLM-NII, WNN-GIP, NetLapRLS, CMF, and BRDTI. The AUC scores of HNEDTI are 0.989, 0.990, 0.965 and 0.913 on Enzyme, Ion Channel, GPCR and Nuclear Receptor, respectively. Thus, we can conclude that HNEDTI can be successfully applied to predict potential drug-target associations.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant No. 61832019, No. 61622213, No.61772552, the 111 Project (No. B18059), the Hunan Provincial Science and Technology Program (2018WK4001), and the Fundamental Research Funds for the Central Universites of Central South University under grant No. 2018zzts560.

References

[1] D. O. Ricke, S. Wang, R. Cai, and D. Cohen, "Genomic approaches to drug discovery," *Current opinion in chemical biology*, vol. 10, no. 4, pp. 303–308, 2006.

[2] T. M. Bakheet and A. J. Doig, "Properties and identification of human protein drug targets," *Bioinformatics*, vol. 25, no. 4, pp. 451–457, 2009.

[3] L.-H. Chu and B.-S. Chen, "Construction of a cancer-perturbed protein-protein interaction network for discovery of apoptosis drug targets," *BMC systems biology*, vol. 2, no. 1, p. 56, 2008.

[4] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, and A. L. Schacht, "How to improve r&d productivity: the pharmaceutical industry's grand challenge," *Nature reviews Drug discovery*, vol. 9, no. 3, p. 203, 2010.

[5] H. Chen and Z. Zhang, "A semi-supervised method for drug-target interaction prediction with consistency in networks," *PloS one*, vol. 8, no. 5, p. e62975, 2013.

[6] T. Pahikkala, A. Airola, S. Pietilä, S. Shakyawar, A. Szwajda, J. Tang, and T. Aittokallio, "Toward more realistic drug–target interaction predictions," *Briefings in bioinformatics*, vol. 16, no. 2, pp. 325–337, 2014.

[7] K. Yang, H. Bai, Q. Ouyang, L. Lai, and C. Tang, "Finding multiple target optimal intervention in disease-related molecular network," *Molecular systems biology*, vol. 4, no. 1, 2008.

[8] X. Chen, M.-X. Liu, and G.-Y. Yan, "Drug–target interaction prediction by random walk on the heterogeneous network," *Molecular BioSystems*, vol. 8, no. 7, pp. 1970–1978, 2012.

[9] F. Cheng, C. Liu, J. Jiang, W. Lu, W. Li, G. Liu, W. Zhou, J. Huang, and Y. Tang, "Prediction of drug-target interactions and drug repositioning via network-based inference," *PLoS computational biology*, vol. 8, no. 5, p. e1002503, 2012.

[10] M. Campillos, M. Kuhn, A.-C. Gavin, L. J. Jensen, and P. Bork, "Drug target identification using side-effect similarity," *Science*, vol. 321, no. 5886, pp. 263–266, 2008.

[11] X. Chen, C. C. Yan, X. Zhang, X. Zhang, F. Dai, J. Yin, and Y. Zhang, "Drug–target interaction prediction: databases, web servers and computational models," *Briefings in bioinformatics*, vol. 17, no. 4, pp. 696–712, 2015.

[12] M. Zeng, F. Zhang, F.-X. Wu, Y. Li, J. Wang, and M. Li, "Protein-protein interaction site prediction through combining local and global features with deep neural networksoriginal paper," *Bioinformatics*, 2019.

[13] M. Zeng, M. Li, Z. Fei, F. Wu, Y. Li, Y. Pan, and J. Wang, "A deep learning framework for identifying essential proteins by integrating multiple types of biological information," *IEEE/ACM transactions on computational biology and bioinformatics*, 2019.

[14] F. Zhang, H. Song, M. Zeng, Y. Li, L. Kurgan, and M. Li, "Deepfunc: A deep learning framework for accurate prediction of protein functions from protein sequences and interactions," *Proteomics*, p. 1900019, 2019.

[15] M. Li, Z. Fei, M. Zeng, F. Wu, Y. Li, Y. Pan, and J. Wang, "Automated icd-9 coding via a deep learning approach," *IEEE/ACM transactions on computational biology and bioinformatics*, 2018.

[16] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa, "Prediction of drug–target interaction networks from the integration of chemical and genomic spaces," *Bioinformatics*, vol. 24, no. 13, pp. i232–i240, 2008.

[17] M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa, "From genomics to chemical genomics: new developments in kegg," *Nucleic acids research*, vol. 34, no. suppl_1, pp. D354–D357, 2006.

[18] I. Schomburg, A. Chang, S. Placzek, C. Söhngen, M. Rother, M. Lang, C. Munaretto, S. Ulas, M. Stelzer, A. Grote *et al.*, "Brenda in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in brenda," *Nucleic acids research*, vol. 41, no. D1, pp. D764–D772, 2012.

[19] N. Hecker, J. Ahmed, J. von Eichborn, M. Dunkel, K. Macha, A. Eckert, M. K. Gilson, P. E. Bourne, and R. Preissner, "Supertarget goes quantitative: update on drug–target interactions," *Nucleic acids research*, vol. 40, no. D1, pp. D1113–D1117, 2011.

[20] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu *et al.*, "Drugbank 4.0: shedding new light on drug metabolism," *Nucleic acids research*, vol. 42, no. D1, pp. D1091–D1097, 2013.

[21] M. Hattori, Y. Okuno, S. Goto, and M. Kanehisa, "Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways," *Journal of the American Chemical Society*, vol. 125, no. 39, pp. 11 853–11 865, 2003.

[22] T. F. Smith, M. S. Waterman *et al.*, "Identification of common molecular subsequences," *Journal of molecular biology*, vol. 147, no. 1, pp. 195–197, 1981.

[23] J.-P. Mei, C.-K. Kwoh, P. Yang, X.-L. Li, and J. Zheng, "Drug–target interaction prediction by learning from local information and neighbors," *Bioinformatics*, vol. 29, no. 2, pp. 238–245, 2012.

[24] T. Van Laarhoven and E. Marchiori, "Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile," *PloS one*, vol. 8, no. 6, p. e66952, 2013.

[25] Z. Xia, L.-Y. Wu, X. Zhou, and S. T. Wong, "Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces," in *BMC systems biology*, vol. 4, no. 2. BioMed Central, 2010, p. S6.

[26] X. Zheng, H. Ding, H. Mamitsuka, and S. Zhu, "Collaborative matrix factorization with multiple similarities for predicting drug-target interactions," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 1025–1033.

[27] L. Peska, K. Buza, and J. Koller, "Drug-target interaction prediction: A bayesian ranking approach," *Computer methods and programs in biomedicine*, vol. 152, pp. 15–21, 2017.